PAPERBACK

ОПТИЧЕСКОЕ УБЕЖИЩЕ ДЛЯ ДАННЫХ

Великолепная тройка наносит ответный удар по энтропии…

Введение

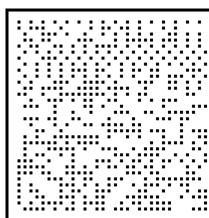
Задача сохранения данных в течение длительного периода времени не относится к числу тривиальных… Размагниченные участки и плотно «спаянные» взаимным притяжением нагромождения выбитых из рабочей поверхности частиц магнитного материала, почти неразличимые участки «ноль» и «единица» на оптическом диске, смешанные до неразличимости уровни flash-ячейки… Поле битвы за данные переместилось из машинных залов в личное пространство… Плотность данных накопителей растет настолько стремительно, что теперь уже не обещанные десятки лет надежного хранения, а даже годы – кажутся пусть и не фантастикой, но уже не гарантированы, и далеко не каждая технология хранения способна обеспечить заявленные 100%...

Назад – в будущее!

Это прозвучит достаточно неожиданно, но существует старая почти как сам мир технология сохранения данных на бумажных носителях. Перфокарты никуда не ушли, а лишь трансформировались в двумерные штрих-коды. Идея хранения данных в виде набора точек нашла применение в виде приложения «PaperBack»: http://habrahabr.ru/post/66004/

Формат A4 позволяет разместить на двух сторонах около 300 Кб данных с приемлемой надежностью (печать при 200 dpi и размере точки 70% от знакоместа, итоговая плотность: ~300 dpi). На практике далеко не каждый бытовой лазерный принтер способен вывести растр с реальной разрешающей способностью более высокой, чем было заявлено выше. Требования к сканеру достаточно приемлемые: требуется около 600 dpi оптического (а не интерполированного) разрешения. «Закатанная» в тонкий «ламинат» бумага, отлично сканируется, не подвержена влиянию перепадов влажности и микроорганизмов… Самое главное – не нарушать герметичность пластикового «бокса» и чрезмерно не царапать поверхность пластика. Положите стопку таких листов в папку, уберите в надежное место – и колода современных «перфокарт» будет служить вам долгие годы. И самое главное – всегда есть возможность визуально убедиться в том, что данные «живы».

Главный герой



«PaperBack» разбивает файл на элементарные блоки размером 32 x 32 пикселя. Полученные 1024 бита (128 байт) разделяются на 96 байт данных (4 байта на адрес блока, 90 байт данных и 2 байта на CRC-16) и 32 байта кодов коррекции ошибок (Рид-Соломон(255,223)). Блоки данных объединяются в группы, разбросанные по странице так, чтобы минимизировать вероятность их одновременного повреждения, и логически связанные через XOR в блок коррекции. Таким образом, если первый слой коррекции (Рид-Соломон) одного из блоков группы не справляется с повреждениями, возможно восстановление целостности на втором уровне (XOR). В противном случае на странице возникает неисправимая ошибка, которая далеко не так страшна, как кажется на первый взгляд. «PaperBack» оснащен технологией «накопления» блоков, и, если одну и ту же страницу распечатать в нескольких копиях и повредить в несовпадающих позициях, при вводе данных «PaperBack» разберется «что к чему». Данные будут собраны в целостную структуру, и порядковый номер страницы исчезнет из списка ожидаемых к сканированию. Кроме блоков, хранящих данные файла, каждая строка страницы содержит по одной копии «суперблок»-а, идентифицирующего файл (имя, время модификации, CRC-16). Таким образом, даже при больших повреждениях страницы, вероятность её достоверной идентификации очень высока. В то же время вероятность появления некорректируемых сбоев в области данных остается далекой от нуля в том случае, когда страницы архива доступны в единичном экземпляре. Для того, чтобы утрата группы блоков, целой страницы архива и даже их набора, не стала преградой на пути считывания данных, требуется применение дополнительного слоя коррекции. Никаких

специальных средств, недоступных рядовому пользователю, не требуется. Но обо всём по порядку…

Великолепная тройка

Список широко распространенного программного обеспечения под Windows, позволяющего снабдить данные кодами для контроля целостности и коррекции ошибок, достаточно краток (если отбросить менее развитые альтернативы): «MultiPar» (Open-source, поддержка форматов семейства «PAR»), «WinRAR5» (проприетарное решение). Оба продукта поддерживают контроль целостности защищаемой структуры данных и способны обнаруживать и корректировать повреждения данных, включающие в себя не только модификацию, но и смещение структуры виртуальных блоков. Совместную работу упомянутых продуктов с «PaperBack» можно сделать очень эффективной, если модифицировать «PaperBack» таким образом, чтобы утилита получила функциональность сохранения неполного набора данных (имеющего частично восстановленные, с ошибками, или утерянные страницы). Кроме того, ошибки в «PaperBack» могут возникнуть и как следствие «пропуска цели», так как компактные сигнатуры целостности блоков (CRC-16) при их большом количестве делают событие «хотя бы один поврежденный блок был признан целостным» вполне вероятным. Общая схема применения проста: перед архивацией на бумагу защитить данные при помощи «MultiPar» или «WinRAR 5», а, затем, без сжатия и шифрования вывести их через «PaperBack». При считывании содержимого архива с бумаги – активировать режим считывания «битых» страниц, и, считав блоки со всех имеющихся бумажных носителей, применить «внешний» слой восстановления данных.

«MultiPar» или «WinRAR 5»?

Практически все схемы коррекции ошибок, применяемые в прикладном ПО, разбивают исходные данные на множество фрагментов одинакового размера – виртуальные блоки. «Стоимость» хранения коррекции ошибок тем выше, чем большее количество виртуальных блоков выделяется в данных, и чем больше самих данных для коррекции предусматривается. Каждый виртуальный блок, который был поврежден, может быть восстановлен, если среди данных для восстановления найдется хотя бы один блок данных для восстановления. В некоторых случаях это равенство может быть нарушено (могут потребоваться дополнительные блоки для восстановления), но в большинстве случаев схема «1:1» работает.

Архиватор «WinRAR5» формирует 200 блоков данных и до 200 блоков для восстановления (при 100% избыточности). Стоит помнить, что любая ошибка, представленная более чем одним измененным битом, может повредить сразу два виртуальных блока, поэтому в худшем случае «WinRAR5» способен откорректировать до 100 многобитовых ошибок. В случае с «MultiPar» это количество может быть значительно увеличено, ценой снижения эффективности кодирования. Снижение эффективности кодирования – это эффект «перетекания» материи «из мышц в кости». Архив с чрезмерным количеством блоков данных можно сравнить с огромным небоскребом, большая часть массы которого – это не полезная нагрузка, а масса стен, которые должны быть достаточно мощными, чтобы выдержать самих себя, именно поэтому «MultiPar» ещё до начала кодирования указывает на текущую эффективность применения выбранной схемы кодирования (количества блоков данных, и блоков для восстановления). Для того, чтобы размер блока «MultiPar» отвечал параметрам физического уровня «PaperBack», требуется устанавливать размер блока на уровне 128 байт. При уровне избыточности в 100% на выходе будет получаться архив, приблизительно равный 400% исходного файла (100% данных и 300% кодов коррекции). Ввиду сниженной из-за малого размера блока данных эффективности, каждые 3% блока данных для восстановления смогут восстановить измененные или смещенные 1% данных в исходном файле.

«WinRAR5» проще в применении и располагает все данные в одном файле, что очень удобно. Эффективность кодирования высока, и там, где пользователь «MultiPar», погнавшись за малым размером блока, распечатает одну копию архива, приверженец «WinRAR5» выведет на печать две копии архивов, имеющих 100% данных для восстановления в своем составе (крупные, относительно плотно сгруппированные ошибки и даже выпадения целых страниц (в пределах внесенной избыточности) для такой схемы не фатальны). С другой стороны, при большом количестве мелких ошибок, схема коррекции с малым размером блока будет иметь более высокую вероятность восстановления. Выбор за вами!